Technology & Standards Watch

Metadata for digital libraries: state of the art and future directions

by

Richard Gartner

Peer Reviewed by:

Hervé L'Hours
Metadata Development & Implementation Manager
Preservation & Systems
UK Data Archive

Grant Young
Project Manager and TASI Technical Research Officer
University of Bristol

To make sure you are reading the latest version of this report, you should always download it from the original source.

Original source	http://www.jisc.ac.uk/techwatch
Version	1.0
First published	April 2008
Publisher	JISC: Bristol, UK

Table of Contents

Executive Summary

At a time when digitization technology has become well established in library operations, the need for a degree of standardization of metadata practices has become more acute, in order to ensure digital libraries the degree of interoperability long established in traditional libraries. The complex metadata requirements of digital objects, which include descriptive, administrative and structural metadata, have so far mitigated against the emergence of a single standard. However, a set of already existing standards, all based on XML architectures, can be combined to produce a coherent, integrated metadata strategy.

An overall framework for a digital object's metadata can be provided by either METS or DIDL, although the wider acceptance of the former within the library community makes it the preferred choice. Descriptive metadata can be handled by either Dublin Core or the more sophisticated MODS standard. Technical metadata, which is contingent on the type of files that make up a digital object, is covered by such standards as MIX (still images), AUDIOMD (audio files), VIDEOMD or PBCORE (video) and TEI Headers (texts). Rights management may be handled by the METS Rights schema or by more complex schemes such as XrML or ODRL. Preservation metadata is best handled by the four schemas that make up the PREMIS standard.

Integrating these standards using the XML namespace mechanism is straightforward technically although some problems can arise with namespaces that are defined with different URIs, or as a result of duplications and consequent redundancies between schemas: these are best resolved by best practice guidelines, several of which are currently under construction.

The next ten years are likely to see further degrees of metadata integration, probably with the consolidation of these multiple standards into a single schema. The digital library community will also work towards firmer standards for metadata content (analogous to AACR2), and software developers will increasingly adopt these standards. The digital library user will benefit from developments in enhanced federated searching and consolidated digital collections. The same developments are likely to take place in the archives and museums sectors, although the different metadata traditions that apply here are likely to make the form they take somewhat different.

The combined benefits of the shared XML platform and the fact that they have already proved themselves in major projects makes these standards the best strategic choices for digital libraries. Although their adoption in integrated environments is still at a relatively early stage, particularly amongst software developers, increasing community-wide use of these will render the production of digital collections easier by freeing resources from metadata to object creation, and facilitate the adoption of service-oriented approaches to core infrastructures. The adoption of integrated metadata strategies should be pressed for at the highest managerial levels.

Keywords: metadata, digital libraries, XML, METS, PREMIS

1. Introduction

Digital library technologies are by now well established and understood throughout the higher education community and the creation of digital collections, either in the form of 'born-digital' materials or the conversion of standard library materials into digital form, is now a well-established part of the activities of most higher education institutions. Making effective use of these resources is dependent on the creation of good quality metadata, without which they cannot be found by users nor administered effectively by their host institutions. To move beyond the ambit of the individual repository so that digital resources can be managed usefully at a sector-wide level, allowing, for instance, collections to be searched together in a federated fashion, requires some degree of standardization of metadata.

This report attempts to provide a snapshot of digital library metadata at a stage when such standardization has become fully practical, even if its possibilities have not yet been fully realized within the higher education sector. The bulk of the report does this by surveying a group of standards which are based on the XML (eXtensible Markup Language) markup language, and showing how they relate to each other in ways which allow them to form an integrated metadata scheme. It will be seen that their application in a scheme of this type is not entirely without its problems, most of which are caused by overlaps and redundancies between the standards, but practical solutions will be outlined for these. An assessment of future developments in library metadata will point to possibilities for consolidating the potential offered by these developments and indicate how they are likely to affect all stakeholders in digital libraries.

N.B. Unless otherwise stated all 'last accessed' dates for weblinks is 9th April 2008.

2. Digital libraries in a networked age

What exactly constitutes a 'digital library' has never been easy to pin down. William Saffady (1995) provided a definition in a seminal article in 1995 that remains pertinent today:-

> "...a library that maintains all, or a substantial part, of its collection in computer-processable form as an alternative, supplement or complement to the conventional printed and microfilm materials that currently dominate library collections." (p. 221)

In the years since Saffady wrote these words, the Internet has become home to digital collections which are wider and more diverse than this rather narrow definition allows: most notably, institutional digital repositories have become important vehicles for the dissemination of research output, museums have increasingly mounted digital collections to complement their curatorial work, and archives have increasingly (at least as far as copyright allows) mounted the collections they hold in their custody for the wider world to access. More recently, with the advent of so-called Web 2.0 technologies (Anderson, 2007) the digital collection has become a more fluid, interactive concept, as applications such as blogs, wikis, and folksonomies become part of the information landscape. While this report concentrates specifically on the core digital library sector, its applicability in principle, at least, to these other sectors will also be considered.

The richness of the collections that have appeared on the Web from these disparate quarters has made an exceptional contribution to the higher education community, but for those hoping to use them the overall experience can be a somewhat daunting one. The overall impression can be of a messy information environment, where every collection has to be searched in different ways through a different interface, and where finding the collections themselves can be, in itself, a difficult and time-consuming process. Even the collections of a single institution created over a period of time may suffer from multiple interfaces and no facilities for cross-searching: Oxford University's digital initiatives[1], for instance, present a diversity of interfaces and disparities of cataloguing standards that result in their usage being more cumbersome and

1 http://www.odl.ox.ac.uk/collections/

time-consuming than the technologies underlying them would potentially allow.

The irony of the ever improving powers of technology failing to deliver their full potential because of an increasingly messy information environment for the user is obvious and needs to be addressed. This is particularly so at a time when the volume of digital collections is likely to increase exponentially now that the importance of the electronic medium for providing access to information and the academic record is fully recognized. To tidy up the information environment, and render it coherent and easily approachable, requires above all an acceptance of the importance of metadata.

3. The need for metadata standardization

Metadata is the core of any information retrieval system and so its implications for any digital library are profound: the choice of a metadata scheme underpins any such library's ability to deliver objects in a meaningful way, and greatly affects its long-term ability to maintain and preserve its digital assets.

The necessity for common approaches to metadata have been acknowledged in the library community for as long as inter-institutional co-operation has been practised. It was recognized particularly in the 1960s when the MARC (Machine Readable Cataloguing) standard and AACR (Anglo-American Cataloguing Rules) cataloguing rules were created to standardize practices into a form which would make full use of the then nascent computing technologies. The MARC standard provided a uniform container to hold cataloguing information in a form that would readily transfer between systems, while AACR provided consistent rules to govern what information would populate the fields of the MARC record and how it would be formatted. It was as a result of the adoption of these standards that the large union catalogues and collaborative cataloguing projects that are now such a prominent part of the library world became possible.

The technology of the digital library offers even greater potential for inter-institutional collaboration: not only can multiple collections be rendered cross-searchable in the style of a union catalogue, but the objects that constitute these collections can themselves readily be integrated into inter-institutional virtual repositories. To do so effectively, however, requires standard approaches to metadata. Without these, problems rapidly arise when digital library collections reach any substantial size: intelligent cross-searching, for instance, becomes very difficult to achieve as inconsistent item descriptions rapidly render retrieval from large collections very imprecise. Without consistency of metadata practices, the often-stated ideal of a 'hybrid library' (Rusbridge, 1998), which integrates traditional and electronic resources, remains a remote possibility.

To adopt an analogy from the traditional library world, it is necessary to standardize both the containers for digital library metadata (a digital library MARC standard) and the rules for the metadata content itself (an analogue to AACR). Unfortunately, adapting these long-established standards for the digital library environment is not a feasible option per se. The metadata required for digital objects is more complicated than that required for physical library items, which is generally limited to describing the intellectual content of an item (such as its author, title or subject), and such basic administrative information (for example, shelfmarks) as is required to curate it. A useful typology for digital library metadata, adopted by early key projects such as the Making of America II (MOA2)[2], indicates the range of information that must be included:--

- **descriptive metadata**: analogous to the tradition catalogue record, this is information on the item's intellectual contents which allows it to be retrieved and its value to the user assessed
- **administrative metadata**: the information necessary to curate the digital item, which includes (not exclusively):-

2 http://sunsite3.berkeley.edu/MOA2/

- technical metadata: all necessary technical information (for example, file formats) to allow the host system to store and process the item
- rights management: declarations of rights held in the item and the information necessary to restrict its delivery to those entitled to access it
- digital provenance: information on the creation and subsequent treatment of the digital item, including details of responsibilities for each event in its lifespan

- **structural metadata**: information necessary to record the internal structure of an item so that it can be rendered to the user in a sensible form (for instance, a book must be delivered in its page order). This type of metadata is necessary as an item may often be comprised of multiple (often thousands) of files - for example, the images of individual pages that make up a digitized book.

Given the complexity of these metadata requirements, it is perhaps not surprising that no single standard has yet emerged which addresses them all. Nonetheless, the emergence of the standards detailed in this report, all of which are based on the Functional Requirements for Bibliographical Records (FRBR[3]) conceptual model, and the interoperability allowed by their common language, does allow for a coherent metadata landscape to be constructed on a sector-wide basis. By adopting the approaches advocated here, it should be possible for the digital library world to achieve a degree of integration that it has formerly lacked, and enhance the possibilities for the types of collaboration elaborated above in ways that have not been possible before.

4. XML: the standard behind the standards

All of the standards discussed here use XML as their semantic and structural underpinning. XML began life in the 1960s as SGML (Standard Generalised Markup Language), a system for tagging up electronic texts using semantically meaningful tags. In addition to marking up texts themselves, it has also come to be used increasingly as a standalone mechanism for encoding metadata for all types of objects in traditional or electronic libraries. There is, for example, a translation to XML of the MARC standard, which encodes the standard MARC fields in XML tags: for example, the title (245) field of the MARC record

```
245 10|aArithmetic /|cCarl Sandburg ; illustrated as an anamorphic
adventure by Ted Rand.
```

is rendered in XML tags as follows:-

```
   <datafield tag="245" ind1="1" ind2="0">
           <subfield code="a">Arithmetic /</subfield>
            <subfield code="c">Carl Sandburg ; illustrated as an
anamorphic adventure by Ted Rand.</subfield>
   </datafield>
```

The strengths of XML as the basis of a metadata scheme have often been acknowledged, for example by UKOLN in their *Good Practice Guide for Developers of Cultural Heritage Web Services: Metadata Sharing*

3 http://www.frbr.org/

and XML (Johnston, 2004). It is a fully open standard registered with the ISO (International Standards Organisation), and so is independent of any given software application. It is acknowledged as the most archivally robust metadata format by, for example, the influential Commission on Preservation and Access (Coleman and Willis, 1997). It also benefits from the relative simplicity of its syntax, combined with great flexibility in the ways in which it can be used: in particular, its ability to encode hierarchical structures by the simple expedient of nesting tags allows it to represent complex relations between metadata components simply and elegantly.

XML applications are usually expressed in what are known as *schemas:* these consist essentially of definitions of the elements which make up the XML tags and of the rules which dictate how they should be used together (for instance, which should nest within which). In the discussion which follows, the term *schema* is used with this technical meaning, in contrast to *scheme* which is used to describe a metadata system as a whole, XML or otherwise.

5. The standards themselves

The basis of any digital library metadata system must be an overarching framework within which everything is held, an equivalent to the MARC record in the traditional library. Two standards have emerged from different communities which aim to fulfil this function: from the library community, specifically the MARC Standards Office who maintain that core cataloguing scheme, comes METS (Metadata Encoding and Transmission Standard[4]), while from the area of multimedia development, in the form of the Moving Pictures Experts Group (MPEG[5]) , comes DIDL (Digital Item Declaration Language[6], also known as part 2 of the MPEG-21 standard).

Both standards attempt to provide overall frameworks within which descriptive, administrative and structural metadata can find logical placings. Both also provide mechanisms for recording inventories of the individual files that make up a digital library object, and methods for recording information on how these should be rendered when delivered to the user (for instance, by specifying what software is necessary to do so). Finally both provide mechanisms for recording the internal structure of a digital object, usually in the form of a nested hierarchy, so that its components can be delivered to the user in a way that makes sense to them.

The ways in which the two standards approach these functions differ in some crucial ways. METS is arranged according to the types of its constituent metadata, each of which (descriptive, administrative and structural) is located in a different section within its overall structure: it then uses a system of pointers to link these sections together so that all the metadata for a given component of the object can be processed as a single entity despite being scattered throughout the file. Providing an overall shape to the digital object is a structural map, a hierarchical representation of the relationships between its components, from which the links to these metadata sections emanate.

Instead of following the METS approach of separating out different metadata types, DIDL collates all types for a given component together: an image file, for instance, will have its descriptive and administrative metadata located together, obviating the need for the complex set of linking mechanisms required by METS. Instead of a separate structural map to act as the centre of a matrix of links to the metadata, DIDL embeds everything into a single hierarchy representing the structure of the digital object.

Both approaches are valid ones and will meet the needs of most digital objects, although the METS approach of separating out metadata by type is undoubtedly the more flexible. Certainly it is the METS standard that has become by far the more widely used of the two, and has established itself as the core metadata

4 http://www.loc.gov/standards/mets
5 http://www.chiariglione.org/mpeg/index.htm
6 http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm

framework in the digital library community: only one major project, the Los Alamos National Laboratory, has made use of DIDL to any major extent (Bekaert et al, 2004). The widespread adoption of METS within the digital library community, and the consequent knowledge base of users, is certainly a strong reason for considering its adoption.

Although METS in some ways aims to be a MARC standard for digital library objects, it differs from MARC in one crucial respect: it does not prescribe any fields for metadata content, instead just defining placeholders for the broad categories (descriptive, administrative, structural) within which metadata is to be logically placed. For the metadata content itself, it is necessary to use other schemes which provide the semantic elements required.

5.1 Descriptive metadata

A number of schemes are available for *descriptive* metadata, of which the two that have established themselves most securely in the digital library world are Dublin Core (DC[7]) and MODS (Metadata Object Description Schema[8]). Dublin Core is perhaps the most widely used scheme for many reasons, of which its simplicity is perhaps the primary: a set of 15 basic fields (such as *creator*, *subject*, *identifier*) designed for resource description and discovery in any type of media, it has formed the basis of many a digital library and underlies further important standards, such as the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH[9]). If these 15 fields prove too broad for a given application, DC offers the option of qualifying them to allow greater precision: the *creator* field, for instance, may be qualified to delineate that role more specifically (such as *creator.author*). While this allows increased precision, it does inevitably reduce the interoperability of DC metadata.

MODS, covered in an earlier TSW report (Gartner, 2003), offers an alternative that effectively gets round this deficiency in DC. It offers a richer set of approximately 80 elements, which allows a much greater degree of precision but retains interoperability by virtue of these elements being fixed and so employed without qualification. Although designed specifically for digital library objects, MODS is based on a subset of the MARC standard and therefore integrates well with metadata held in traditional library catalogues. It also incorporates the facility to extend its element set in the very rare cases where this is needed (although this will, of course, reduce its interoperability slightly).

Both DC and MODS have committed adherents in the digital library sector, and it is difficult at present to imagine either becoming predominant. The wide constituency of DC users, and the increasing set of community-agreed extensions to it (with their concomitant support bases), both make it a safe option for the digital library designer, despite its limitations stated above. The advantages of MODS, which, despite its origin as a derivative from MARC, is not merely a bibliographic standard, merit serious consideration, particularly for metadata requirements of any complexity. Either, however, are perfectly sensible options and the choice of which is adopted may often depend on the existing expertise of implementers.

5.2 Administrative metadata

Technical metadata

The choice of standards for technical metadata will inevitably depend on on the type of files that make up the digital object. Generally, though not always, there is one well-established standard for each type of file.

7 http://dublincore.org/
8 http://www.loc.gov/standards/mods/
9 http://www.openarchives.org/OAI/openarchivesprotocol.html

Still images

For still images, the generally accepted standard is MIX (Metadata for Images in XML[10]). MIX is essentially a version in XML of an extensive set of elements devised by the National Information Standards Organisation (NISO) for the detailed technical description of still images. The range of information that can be encoded in MIX is very large, from basic information on file types and sizes, to details of image capture (including capture hardware and image targets), to details of how an image has been processed after capture. Although a MIX file can be very lengthy and complex, almost all of its components (more than in its parent element set) are optional so that a basic record may be very simple. Although still in the process of revision (version 2.0 is in draft at the time of writing), MIX has already established itself as the key standard for this type of metadata.

Text

An electronic text may at first sight appear to have minimal technical metadata requirements, but certain key features, such as its character set, languages and byte order, do require documentation to ensure its viability as an electronic object. One possibility is to use standalone 'headers' encoded in TEI (Text Encoding Initiative[11]) format: these are used to record the metadata associated with a TEI document and can be held separately from the texts themselves. They contain a very rich set of elements for describing features of the encoded text in great detail and so represent a viable option where this level is required (for instance, when the textual objects are to form the basis of analytical processing).

A less comprehensive, but simpler, option that is adequate for textual objects that do not require the level of detail provided by the TEI header is the *Schema for Technical Metadata for Text*[12] created by New York University; this is a set of 16 elements that provides all the information necessary for the rendering and display of texts in most contexts.

Audio

For the technical metadata associated with audio files, nothing as predominant as the MIX schema for still images has yet been produced. There is, however, a very useful schema produced by the Library of Congress for its digital library projects called AUDIOMD (Audio Technical Metadata Extension Schema[13]). The elements provided by this schema include all key information necessary to make sense of the audio file (including, for instance, its format, bit rates, sampling frequencies, and any compression applied to it). This is a very usable schema of a minimum size and complexity necessary to be functional, that fits well into the metadata landscape.

Video

As with audio, no single schema for video files has yet emerged to be a predominant choice for this type of object. An important schema from the digital library world itself is the Library of Congress' *Video Technical Metadata Extension Schema* (VIDEOMD[14]). This set of 36 elements has been designed for the Library's own digital library projects and fits in very well with the requirements of video in the digital repository. It concentrates solely on technical metadata and so avoids any potential problems of overlap with schemes for other types of metadata.

10 http://www.loc.gov/standards/mix/
11 http://www.tei-c.org/
12 http://dlib.nyu.edu/METS/textmd.htm
13 http://lcweb2.loc.gov/mets/Schemas/AMD.xsd
14 http://lcweb2.loc.gov/mets/Schemas/VMD.xsd

Another schema that has established itself as a popular choice is PBCore[15], produced by public broadcast television services in the USA. This includes a comprehensive set of elements for the technical description of the digital video file itself, including details of file formats, encoding, duration, aspect ratios and details of changes made as it is processed after creation. It also includes elements for the file' s descriptive metadata and intellectual property information, which can cause problems by overlapping with other schemas such as MODS or PREMIS.

Rights Metadata

Managing access to the digital object is a key function of any digital library system and so a scheme to codify the rights inherent in an object, and to specify to whom and on what conditions it should be made available, is a key part of any metadata environment. One simple scheme devised specifically for use with METS is the *METS Schema for Rights Declaration*[16], which includes a declaration of the type of rights held, contact information for the rights holders and a section detailing by whom and in what circumstances the object may be accessed.

Other, more detailed, schemas are currently being promoted from within the commercial sector: XrML (eXtensible Rights Markup Language[17]), for example, was originally developed by Microsoft and Xerox, but has since become an official part of the MPEG-21 standard. This has now reached a 'fixed' form in version 2.0 and can now be considered a serious option. ODRL (Open Digital Rights Language[18]) is an alternative from the open-access movement, released under the Creative Commons agreement. This is in more of a state of flux than XrML, although a definitive version (2.0) is expected in the near future. Both of these are comprehensive standards that will integrate well with digital rights management systems, and should now be considered serious options for the developer: the open-access model of ODRL may make it more appealing to the higher education sector, although XrML is free to use if not done in conjunction with some of the patented technologies designed around it. Further information on these standards is available is another TSW report (Barlas, 2006).

Preservation metadata

Beyond the immediate concerns of description and technical and administrative management, the effective long-term preservation of a digital object requires further metadata specific to this function. The type of information that needs to be recorded include details of provenance and ownership, 'fixity' information that is required to test the authenticity and validity of the item's constituent files, an event log to record actions performed on it, and any technical and rights information that is necessary to deliver it to the end user. Clearly there is some overlap, particularly in terms of technical and rights metadata, with other schemas, although much of the information that comes under the umbrella of preservation metadata is particular to it.

The best-established schema to deal with preservation metadata is undoubtedly PREMIS (PREservation Metadata Implementation Strategies[19]), the product of extensive work by an authoritative, international working group. PREMIS is essentially a data dictionary, a set of elements from which a number of separate XML schemas have been derived: these cover respectively:-

- the *object* itself, including identifiers, checksums, information on its creation and its relationships to other objects
- *events* associated with it, such as its creation and how and when it has been processed thereafter
- *agents* associated with its preservation (people, organizations and software)

15 http://www.pbcore.org/
16 http://cosimo.stanford.edu/sdr/metsrights.xsd
17 http://www.xrml.org/
18 http://odrl.net/
19 http://www.oclc.org/research/projects/pmwg/

- *rights* associated with it.

In addition to these separate schemas, an overarching framework schema is also available into which all of these can be slotted.

It will be noticed immediately that there is considerable overlap between the contents of these schemas and some of the others already discussed: how some of these duplications may be resolved will be discussed in a later section.

6. Producing an integrated metadata landscape

Although none of these schemas on its own provides all of the elements needed to meet the requirements of a comprehensive digital library metadata environment, taken together they do form a viable system of this kind. To do so, they must be integrated in a simple and reliable fashion, which is made possible by the availability of all as XML schemas (although this may not by any means be their only manifestation). XML has the crucial feature that a marked-up file can embed others encoded in different XML schemas directly within it (if, of course, it follows a schema that is designed with this function in its specification). This is made possible by a feature known as XML *namespaces*.

Namespaces are a mechanism for ensuring that otherwise identical element names can be differentiated from each other. This is done by defining a short prefix to precede element names which is declared by associating it with a URI (Uniform Resource Identifier): for example, the namespace *xhtml* is declared as follows:-

```
xmlns:xhtml="http://www.w3.org/1999/xhtml"
```

Although this looks like an online address, the URI given is used essentially as a lengthy string to differentiate elements with different namespace prefixes. So, for instance, the element

```
xhtml:title
```

is expanded by an XML process to

```
http://www.w3.org/1999/xhtml:title
```

to differentiate it from any other *title* elements within a document.

Namespaces are a crucial feature of the METS standard, as they allow metadata encoded in any XML schema to be incorporated into METS files. In a METS application, the subsidiary schemas used (often known as *extension schemas*), and the namespace prefixes by which their elements are to be identified, are declared at the beginning of the METS document: for example, if MODS records are to be used for descriptive metadata, MODS is declared as follows:-

```
xmlns:mods="http://www.loc.gov/mods/v3"
```

and a further declaration points the XML processor to the location of the MODS schema file itself for validation purposes:-

```
xsi:schemaLocation="http://www.loc.gov/mods/v3/schemasdirectory/mods-3-0.xsd"
```

As mentioned earlier, the METS file is divided into clearly delineated sections each containing a different type of metadata. Using the schemas discussed earlier and the namespacing mechanism outlined above, it is possible to populate the METS framework in a straightforward way.

6.1 Descriptive metadata

Both MODS and the DC schema fit cleanly into METS's descriptive metadata section, delineated by their respective namespaces. METS allows multiple descriptive metadata sections to meet any requirements that cannot be met by a single record alone: it is possible, for instance, to include more than one instance of MODS to accommodate descriptions at different levels of detail or in multiple languages.

6.2 Administrative metadata

METS divides its administrative metadata section into several smaller subsections, each of which can be populated with one or other of the schemas listed above. Specifically:

- **technical metadata**: the schema used here will obviously depend on the type of files included in the object; more than one type of file will require the use of more than one type of metadata schema. The most obvious choices are:-
  - still images: MIX
  - video: VIDEOMD (or possibly PBCore)
  - audio: AUDIOMD
  - text: *Schema for Technical Metadata for Text* or TEI Headers
- **rights metadata**: the *METS Schema for Rights Declaration* for straightforward ownership declarations and access control, or XrML/ODRL for more complex rights control. The *rights* schema of PREMIS also fits in here.
- **source metadata**: this is metadata about the original item from which the digital copy was made – it will, of course, be absent if the item is *born digital*. This is predominantly descriptive metadata, and so will generally use MODS or DC (although, of course, more specialized schemas may be used where required).
- **digital provenance metadata**: this is essentially an audit trail for the digital object, tracing its creation and the changes that have been made to it. The *events* schema of PREMIS fulfils this function and should be located here.

This arrangement accounts for all the metadata schemes outlined above apart from PREMIS's *object* and *agent* schemas. Its *object* schema generally (but not entirely) covers technical information for an item and so should be placed in METS's technical metadata section. The *agent* schema is not so clear-cut: if the agent is associated with some event in the item's creation or subsequent processing it should go into the digital

provenance section, but if it is associated with a permission for usage it should clearly fit into rights metadata.

6.3 Structural metadata

As mentioned earlier, METS provides a hierarchical structural map to encode metadata on the internal structure of an item. This is simply a series of nested elements, named *div* (short for division) whose nesting is meant to emulate the structure of the digital object: so, for example, a digitized book would have its structure of *divs* arranged to match its original divisions into chapters, sections of chapters and so on. The structural map may encode either a physical or a logical structure: it may, for instance, describe the pagination structure of a volume, or the arrangement of its intellectual contents. Often these will coincide, but where this is not the case it is entirely feasible to include multiple structural maps, each describing a different type of structure, and to link them together by METS's comprehensive linking facilities.

7. Some problems with this approach

From the above, it is clear that METS and the schemas detailed above provide an integrated XML metadata environment for digital libraries. However, using such a heterogeneous range of schemes which have not necessarily been constructed with this overall environment in mind can cause a few problems: in almost all cases, these can be resolved by the pragmatic application of good practice guidelines.

One minor technical problem that can arise is that the XML schemas used within the METS framework may themselves incorporate subsidiary schemas whose namespace definitions may conflict with those in METS, or with others embedded into the same METS file. This problem arose for a while with the *xlink* schema, a set of elements used to provide linking mechanisms[20]: both METS and MODS used this schema for this purpose, but each pointed to a different URI when defining it; this was subsequently corrected by a modification to the METS schema. Theoretically, this should not cause problems as the first definition encountered by an XML parser (which in this case would be that in the METS file) should take precedence, but in practice, several well-known XML software packages cannot resolve this conflict and will not validate files in which it occurs; it remains a potential problem in any environment where multiple schemas are used in conjunction.

More problematic are circumstances in which schemas do not fit cleanly into their respective slots within the METS framework. The most obvious example of this is PREMIS: as has already been seen, although three of its schemas fit readily into sections of the METS file (*object* into technical metadata, *event* into digital provenance, and *rights* into rights metadata), its *agent* schema is ambiguous within the METS scheme and could go into either the digital provenance or rights sections according to its function. The problem is exacerbated if it is desired to keep all PREMIS metadata together: its *container* schema could fit logically under either technical or digital provenance metadata, but this is an untidy solution which inevitably involves some of its elements being left in an illogical location.

Beyond the issues of what should go where, redundancies between metadata schemas are certain to occur when they are combined within the METS framework. There is, for example, a considerable duplication of elements between PREMIS's *object* schema and MIX (such as information on checksums for file validation). Similarly there are many redundancies between PREMIS and the METS schema itself, particularly in the area of structural metadata where PREMIS's elements that cover relationships between components can clash with the METS structural map. Decisions have to be made on whether to accept such redundancies and

20 XLink is a vocabulary that allows hyperlinking to be added to any XML document. In order to use it, XLink-namespaced attributes are added to schema elements. See: http://www.w3.org/TR/xlink/

duplicate information in both sections, or whether to include such metadata in one only. If information is to be duplicated, a set of rules must be drawn up to ensure that the processor knows which manifestation is to be given priority.

Issues of this type are currently being addressed by the library community and good practice guidelines are being drawn up. A set of best practices dealing specifically with the problems arising from the use of PREMIS and METS, for instance, are (at the time of writing) in an advanced draft stage by the Library of Congress (Guenther, 2007) . Clearly, following any such best practices as they are defined is the most responsible course of action. Until they are fully established, it is important that projects document the decisions they have made to handle these issues: in the case of METS-based projects, this can be done most effectively by the use of METS profiles[21], XML documents which record the technical details of a given application of METS and are registered with a central repository at the Library of Congress.

8. Likely future developments

The next five to ten years will undoubtedly see moves towards the further integration of metadata standards that will consolidate trends that are already manifest. The XML platform shared by these schemas already ensures a degree of interoperability that allows them to be used together to form a integrated metadata landscape. The most likely future trend stemming from this will be their consolidation further into a single schema. The benefits accruing from such a single, off-the-shelf, schema would be considerable, above all by greatly reducing the learning curves associated with the use of the current range of standards and lowering the barriers to adoption that arise from these. Such a solution would be preferable to the use of disparate schemas, despite the flexibility allowed by the namespacing mechanism which allows them to function as a single entity.

A concurrent strand of developments is likely to take place in the form of the compilation of agreed standards for metadata *content*. At present there is little co-ordination of approaches to mirror the extensive range of cataloguing rules that apply (in the form of AACR2) in traditional library settings. Few digital library sites, for instance, follow the name authority conventions which in traditional catalogues cover personal, geographic and corporate names and have led to the construction of large authority files such as the 10 million strong Library of Congress Name Authority File (LCNAF[22]). This state of affairs is not viable in the long run, particularly in an age of federated searching where the lack of a standard form of description can severely reduce the precision and recall of search results. Moves to rectify this are already well under way in the form of Resource Description and Access (RDA[23]), a set of proposals to replace AACR2 with new rules for bibliographic description and authority controls which is due for publication in 2009: these guidelines, if adopted, should extend the standardization offered by AACR2 to a much wider range of materials, in particular those in digital formats. To accompany this, it will also be necessary for community-wide efforts to be made to extend current name authority files, which are generally based on persons associated with published works, to incorporate the wider range of names associated with digital objects.

An essential complement to the development of these standards will be their increasing adoption by digital library software developers: the clear advantages of open, integrated metadata standards will inevitably mean that software that uses proprietary metadata schemes will become less viable. The adoption of these standards is, in fact, already under way in both the open-source and commercial sectors: in the former, for instance, both Fedora[24] and Greenstone[25] currently use METS whilst among commercial products,VTLS's

21 http://www.loc.gov/standards/mets/mets-profiles.html
22 http://authorities.loc.gov/
23 http://www.collectionscanada.gc.ca/jsc/rda.html
24 http://www.fedora.info/
25 http://www.greenstone.org/

VITAL[26], based on Fedora, is also METS-compatible. This trend is certain to continue: indeed, it is difficult to imagine that in 10 years any system not conforming to these standards would be any more viable than a library management system would be today without MARC compatibility.

For the library user, the most obvious manifestation of the adoption of these standards will be the appearance of more advanced federated facilities that will allow disparate digital library collections to become co-searchable. Some examples of the possibilities that lie ahead can already be seen in still experimental services such as OAIster[27], which provides basic cross-searching of over 14 million resources encoded in simple Dublin Core. Future services, based on the much richer possibilities presented by more sophisticated schemes such as MODS, are likely to offer more advanced cross-searching facilities.

Outside the digital library sector itself, many of these trends are also likely to manifest themselves, although the different metadata traditions within the archives community (based on ISAD-G[28]) or the museum sector (which generally follows the CIDOC Conceptual Reference Model[29]) make it unlikely that these will adopt the standards discussed above, which are based essentially on the FRBR model. Despite these differences, there has already been a consolidation of metadata practices in the archives sector (in the form of EAD[30]) and in museums (with the widespread adoption of the SPECTRUM[31] standard), indicating that the imperatives of metadata standardization are well recognised within them. Software developers have increasingly adopted these standards, and users have already benefitted from consolidated services (such as Access to Archives[32]) which exploit the possibilities of a unified metadata strategy. Although, therefore, the form taken by these developments is likely to be different in these cases, the overall principles and imperatives of a co-ordinated metadata strategy at the sector-wide level will remain just as pertinent.

9. Assessment: why these standards matter

The need for integrated metadata strategies within the digital library environment is now well recognised, and so the utility of these standards will to a great extent depend on their ability to provide this level of integration. The common platform of XML is of great importance to this strategy: without the ability of XML to allow the type of embedding demonstrated above, and the use of namespaces to allow this to operate in a manageable way, this approach would be unlikely to succeed. It is this facility, which allows the flexibility of being able to choose the most relevant standards while maintaining a single architecture, that makes such a integrated strategy using separately constructed schemes feasible. Despite its usefulness, however, it should only be seen as an interim solution towards fully integrated standards as envisaged in the last section. Combining metadata standards, even a limited set such as described above, will always be messier than utilising a single standard that combines their taxonomic powers and resolves any potential clashes or duplications between them.

Integration by itself would, of course, be of little consequence if the standards themselves failed to address the metadata needs of the digital library community. In this respect, the provenance of each standard is of some importance. All have been constructed by authoritative standards setters within their communities: METS and MODS, for instance, are maintained by the MARC Standards Office which also administers the profession's core library cataloguing standards, MIX by NISO, and PREMIS by a large number of authoritative bodies in the area of metadata. Acceptance by the library community is vital if these standards

26 http://www.vtls.com/products/vital
27 http://www.oaister.org/
28 http://www.ica.org/en/node/30000
29 http://cidoc.ics.forth.gr/index.html
30 http://www.loc.gov/ead/
31 http://www.mda.org.uk/spectrum.htm
32 http://www.a2a.org.uk/

are to achieve the critical mass necessary to embed themselves in the digital library world: their respective provenances certainly make this more likely.

Beyond these questions of provenance, all of these standards have proven themselves in complex practical applications. METS, for instance, has been adopted by many key players in the digital library world including the Library of Congress, and has proved its ability to meet the requirements of major and highly complex digital collections. MODS has now such a widely accepted user base that its viability is beyond doubt. None of these standards are therefore unproven or likely to cause any problems of usability. Despite the substantial advantages accruing from the adoption of these standards, a few problems arise from the fact that this is a relatively early stage in the development of digital library metadata strategies. There is, for instance, still an occasionally limited awareness of these standards and how they may be used together within the digital library community, which will require education and institutional support to alleviate. Similarly, awareness amongst software developers is still at a relatively early stage and many important applications do not currently use them: this limits the range of platforms available to digital library managers who wish to adopt them. This situation will certainly improve as the digital library community asserts itself more vigorously in its dealings with vendors and obligates them to adopt more integrated approaches based on open standards.

Beyond these, perhaps, ephemeral considerations which reflect a given stage in the lifecycle of the digital library, the importance of this approach extends into the overall philosophy that should underlie a coherent and robust approach to metadata. The vision of integrated and service-oriented approaches to digital libraries is becoming an increasingly common one and integrated approaches to metadata will necessarily be part of their realization. The European DELOS Association's Reference Model for Digital Library Management Systems[33] for instance, which advocates a unified model to integrate users with content and architecture (Candela, 2007), will require a coherent approach to metadata to link these components into a single conceptual entity. Similarly, the service-oriented approaches towards interoperability proposed by JISC's e-Framework[34], which aims to foster the development of small, easily combined components which readily exchange data (Olivier, 2007), would clearly only be viable with the standardization of metadata to facilitate this.

Clearly, the adoption of an open set of standards linked in a clearly-established manner will ultimately make the process of creating digital libraries much easier for all concerned: the waste of duplicated effort caused by the 'reinvention of the wheel' that too often still occurs as each project devises its own metadata scheme can be greatly reduced if a strategy of the type outlined here is adopted throughout the community. This would leave more energies, in terms of both personnel and finance, to be devoted to the creation and management of the digital collections themselves.

Much of the direction that metadata in the digital library community, defined in its widest sense, will take in the next five to ten years will inevitably depend on the requirements, some technical, others political, of the major decision-making bodies in that area. Certainly a recognition by such bodies of the value of integrated metadata, particularly for the possibilities it allows for the re-use of the corpus of materials created as a result of the projects that they facilitate, is important and should be pressed for. It is in providing the 'glue' that would underlie the service orientation of future directions that an integrated metadata strategy would have its most profound impact, and it is for that reason that its adoption should be argued for at the highest managerial levels.

33 http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid=
34 http://e-framework.org/

Glossary

AACR	Anglo-American Cataloguing Rules
AACR2	Anglo-American Cataloguing Rules, version 2
CIDOC	International Committee for Museum Documentation
DC	Dublin Core
DIDL	Digital Item Declaration Language
DLF	Digital Library Federation
FRBR	Fundamental Requirements for Bibliographic Records
ISAD-G	International Standard Archival Description (General)
ISO	International Standards Organisation
JISC	Joint Information Systems Committee
LCNAF	Library of Congress Name Authority File
MARC	MAchine Readable Cataloguing
METS	Metadata Encoding and Transmission Standard
MIX	Metadata for Images in XML
MODS	Metadata Object Description Schema
MPEG	Moving Pictures Experts Group
NISO	National Information Standards Organisation
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ODRL	Open Digital Rights Language
PREMIS	PREservation Metadata Implementation Strategies
SGML	Standard Generalised Markup Language
TEI	Text Encoding Initiative
TSW	Technology and Standards Watch (JISC)
URI	Uniform Resource Identifier
VIDEOMD	Video Technical Metadata Schema
VTLS	Virginia Tech Library Systems
XML	eXtensible Markup Language

References

ANDERSON, P. 2007. *What is Web 2.0? Ideas, technologies and implications for education.* Published by JISC Technology and Standards Watch. Available online at: http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf [last accessed: 9th April 2008]

BARLAS, C. 2006. *Digital Rights Expression Languages (DRELs)*. JISC Technology & Standards Watch, July 2006. Available online at: http://www.jisc.ac.uk/whatwedo/services/services_techwatch/techwatch/techwatch_ic_reports2005_published.aspx [last accessed: 9th April 2008]

BEKAERT, J., HOCHSTENBACH, P., VAN DE SOMPEL, H. 2003. *Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library.* **D-Lib Magazine**, Volume 9, Number 11, November 2003. Available online at: http://www.dlib.org/dlib/november03/bekaert/11bekaert.html [last accessed: 9th April 2008]

CANDELA, L., CASTELLI, D., PAGANO, P. 2007. *A Reference Architecture for Digital Library Systems: Principles and Applications*. **Lecture Notes in Computer Science**, Springer Berlin. Volume 4877/2007. DOI 10.1007/978-3-540-77088-6

COLEMAN, J., WILLIS, D. 1997. *SGML as a Framework for Digital Preservation and Access.* Commission on Preservation and Access. Available online at: http://eric.ed.gov:80/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED417748&ERICExtSearch_SearchType_0=no&accno=ED417748 [last accessed: 9th April 2008]

GARTNER, R. 2003. MODS: Metadata Object Description Schema. JISC Technology & Standards Watch. October 2003. Availlable online at: http://www.jisc.ac.uk/whatwedo/services/services_techwatch/techwatch/techwatch_report_0306.aspx [last accessed: 9th April 2008]

GUENTHER, R. 2007. Best practices for using PREMIS with METS. The Library of Congress (Draft), 9th August 2007. Available online at: http://www.loc.gov/standards/premis/best-practices-premismets-20070809.doc [last accessed: 9th April 2008]

JOHNSTON, P. 2004. *Good Practice Guide for Developers of Cultural Heritage Web Services*. UKOLN. Available online at: http://www.ukoln.ac.uk/interop-focus/gpg/Metadata/#section2 [last accessed: 9th April 2008]

OLIVIER, B. 2007. *Having your cake and eating it: The e-Framework's Service-Oriented Approach to IT in Higher Education.* **Educause Review**. vol. 42, no. 4 (July/August 2007), pp. 58–67. Available online at:http://connect.educause.edu/Library/EDUCAUSE+Review/HavingYourCakeandEatingIt/44596?time=1207577197 [last accessed: 9th April 2008]

RUSBRIDGE, C. 1998. *Towards the Hybrid Library.* **D-Lib Magazine**. July/August 1998. Available online at: http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html [last accessed: 9th April 2008]

SAFFADY, W. 1995. *Digital library concepts and technologies for the management of library collections: an analysis of methods and costs.* **Library Technology Reports,**. Vol. 31, No. 3, pp. 221-380. Available for purchase online at: http://cat.inist.fr/?aModele=afficheN&cpsidt=2485292

About the author

Richard Gartner is an information professional who has specialized in the fields of electronic information provision for over 20 years. From 1991-2007 he was New Media Librarian for Oxford University Libraries, where he was responsible for the introduction of the Internet into the Bodleian Library, the Library's first CD-ROM network and its first digital imaging projects. In recent years, he has specialized in metadata for digital libraries, in which capacity he is a member of the editorial board for the METS (Metadata Encoding and Transmission Standard) standard for digital library metadata.

Richard can be contacted by email at: richardoxford@gmail.com